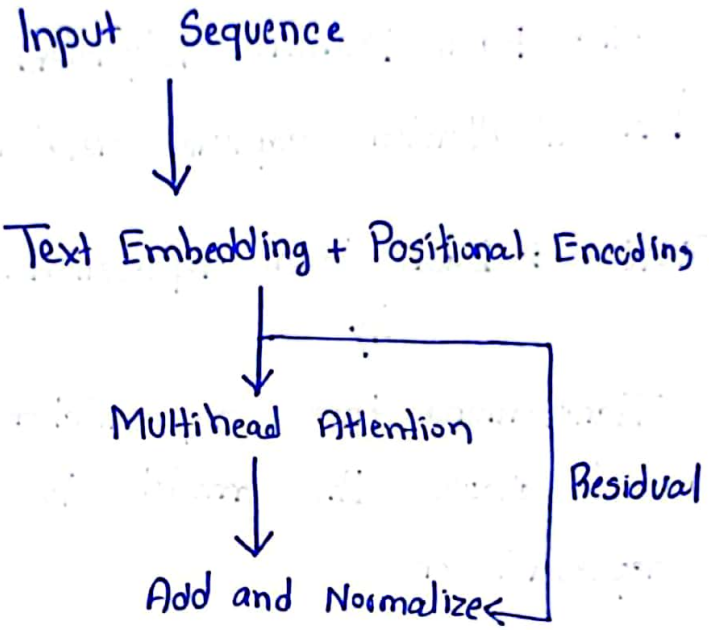
## → Complete Encoder Transformer Architecture




dimensions

$Q = 64$

$K = 64$

$V = 64$

$\sqrt{d_k} = \sqrt{64}$

$= 8$

$P.E = 512$

Multihead attention

$= [z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8]$

### > Residual Connection

↳ Skip connection

① Addressing the Vanishing Gradient Problem

↳ Residual connections create a short path for gradients to flow directly through the networks. Gradients remains sufficiently large.

② Improve gradient flow

↳ Convergence will be faster

③ Enables training of Deeper Networks

## > Feed Forward Neural Network

① Add Non-Linearity

② Processing Each Position Independently

↳ Self attention captures relationship

FFNN ⟶ Each token representation independently

⇓

Transforming these representations furthers and allows the model to learn richer representation

③ FFNN ⟶ Deeper

↳ Add depths to the model

Depth ↑↑ ⇒ More learnings ⟶ DATA